doi:10.1016/j.jmb.2008.08.038 J. Mol. Biol. (2008) 383, 731–744

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Characterization of Protein Fold by Wide-Angle X-ray Solution Scattering

Lee Makowski*, Diane J. Rodi, Suneeta Mandava, Satish Devarapalli and Robert F. Fischetti

Biosciences Division,
Argonne National Laboratory,
9700 South Cass Avenue,
Argonne, IL 60439, USA

Received 8 July 2008;
received in revised form
13 August 2008;
accepted 14 August 2008
Available online
23 August 2008

Edited by I. Wilson

Wide-angle X-ray solution scattering (WAXS) patterns contain substantial information about the three-dimensional structure of a protein. Although WAXS data have far less information than is required for determination of a full three-dimensional structure, the actual amount of information contained in a WAXS pattern has not been carefully quantified. Here we carry out an analysis of the amount of information that can be extracted from a WAXS pattern and demonstrate that it is adequate to estimate the secondary-structure content of a protein and to strongly limit its possible tertiary structures. WAXS patterns computed from the atomic coordinates of a set of 498 protein domains representing all of known fold space were used as the basis for constructing a multidimensional space of all corresponding WAXS patterns ('WAXS space'). Within WAXS space, each scattering pattern is represented by a single vector. A principal components analysis was carried out to identify those directions in WAXS space that provide the greatest discrimination among patterns. The number of dimensions that provide significant discrimination among protein folds agrees well with the number of independent parameters estimated from a naïve Shannon sampling theorem approach. Estimates of the relative abundances of secondary structures were made using training/test sets derived from this data set. The average error in the estimate of α-helical content was 11%, and of β-sheet content was 9%. The distribution of proteins that are members of the four structure classes, α, β, α/β and α+β, are well separated in WAXS space when data extending to a spacing of 2.2 Å are used. Quantification of the information embedded within a WAXS pattern indicates that these data can be used as a powerful constraint in homology modeling of protein structures.

Keywords: WAXS; protein fold; secondary structure; tertiary structure

Introduction

Short of a full, three-dimensional structure determination there are few biophysical techniques that provide information adequate for inference of protein fold. Wide-angle X-ray solution scattering (WAXS) is one technique with that potential. WAXS has proven to be highly sensitive to small changes in secondary, tertiary and quaternary structure[1–5] when scattering of X-rays from a solution of proteins is measured to angles comparable to that used in crystallography ($d \sim 2$ Å). Because of the random tumbling of molecules diffusing in the irradiated sample, the solution scattering pattern is circularly symmetric and can be averaged into a one-dimensional distribution of intensity as a function of scattering angle. Compared to the three-dimensional data sets obtained in X-ray crystallography, the information in a WAXS pattern is modest, providing information adequate for the calculation of a pair-distribution function—literally, a histogram of all the interatomic vector lengths in the protein. Nevertheless, distribution of interatomic vector lengths represents a valuable set of information for determining the secondary and tertiary structure of a

*Corresponding author. E-mail address: lmakowski@anl.gov.

Abbreviations used: WAXS, wide-angle X-ray solution scattering; SAXS, small-angle X-ray scattering; PCA, principal components analysis; TIM, triose phosphate isomerase.

protein. Secondary and tertiary structures exhibit periodicities that give rise to characteristic patterns of interatomic vector lengths, suggesting that the type of information available from a WAXS pattern may be well suited for constructing a restricted set of possible structures of a protein. However, it has been suggested that WAXS, by itself, provides little information about the secondary-structure content of a polypeptide or its side-chain packing.[6] Consequently, a quantitative analysis of the potential utility of WAXS for distinguishing folds is needed. Here we demonstrate that the information present in a WAXS pattern is not only sufficient to provide a good estimate of the abundances of secondary structures, but also to provide a strong restriction on the protein's tertiary structure.

The fold of a protein is encoded in the map of all interatomic vectors of the protein. The software package DALI, which is routinely used for comparison of the three-dimensional structures of proteins,[7] relies on comparison of distance matrices that contain all the C^{α}–C^{α} distances in the protein. The distance matrices used in these calculations contain (except for overall chirality) all the information needed to reconstruct the full three-dimensional structure of a protein.[7] Extensive use of DALI has demonstrated the power of distance matrices in identifying proteins of similar structure and defining the distribution of proteins in 'fold space'[8] or 'protein structure space'.[9] Proteins with similar structures cluster in this fold space. Proteins sharing similar function also colocalize in this space.[9] The information embedded in a WAXS pattern is similar in kind to that of a distance matrix in that the patterns are a reflection of all of the interatomic vector *lengths* in the protein. Here we approach the problem of quantifying the potential for using WAXS patterns to characterize protein structures by constructing a 'space' of the WAXS patterns corresponding to all known protein folds and determining whether colocalization of proteins in this space indicates similarity of folds and possibly functions.

A WAXS space was constructed using a population of computationally generated WAXS patterns derived from the atomic coordinates of 498 protein domains that represent known fold-space as defined by SCOP.[8,10] Analysis of the distribution of patterns in this space provides an estimate of the number of independent parameters in a solution scattering pattern that is in agreement with Shannon sampling theory. We show that data extending to ~10 Å spacing ($1/d=s<0.1$ Å^{-1}) provide little or no information about protein fold, whereas data extending from 10 to 2.2 Å spacing ($0.1<s<0.45$ Å^{-1}) contain significant information pertaining to protein fold. Both size and shape of the protein as well as distinctions among the major protein classes (α, α/β, α+β and β) are represented in the most significant eigenvectors characterizing the set of computed WAXS patterns. Analysis of clustering within WAXS space indicates that comparison of a WAXS pattern from a protein of unknown structure with patterns from proteins of known structure is a viable technique for compiling a short list of possible fold assignments for a protein of unknown structure.

The WAXS patterns analyzed here were computationally generated using the program CRYSOL,[11] the most widely used program for the calculation of small-angle X-ray scattering (SAXS)/WAXS patterns from atomic coordinates. CRYSOL uses a multipole expansion for fast calculation of the spherically averaged scattering pattern and takes into account the hydration shell and the excluded volume. The effect of the exclusion of water from the volume occupied by the protein is approximated with the form factors introduced by Fraser *et al*.[12] These form factors place at the position of each atom a negative Gaussian of weight and volume representative of the water displaced by the atom. CRYSOL, originally designed for calculation of SAXS patterns, has some shortcomings when used to calculate wide-angle patterns from large proteins. However, as shown in Fig. 1, it has proven reasonably accurate

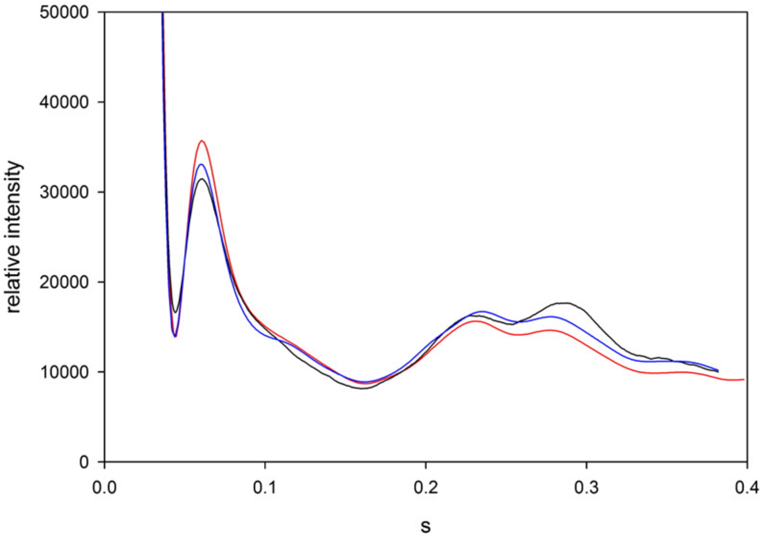

Fig. 1. Plot comparing the experimental WAXS pattern[1] from cytochrome *c* (black); with that computed using CRYSOL (blue); and that calculated using the Debye formula [Eq. (1)] with atomic scattering factors and dummy atoms to represent excluded volume as introduced by Fraser *et al.*[12] (red). CRYSOL reproduces the major features of the observed pattern. The differences between CRYSOL and the Debye calculation are due to (i) the fact that CRYSOL takes into account the hydration layer and (ii) the multipole expansion used by CRYSOL. The difference between observed and calculated is due largely to small intramolecular motions of the protein in solution.[13]

for calculation of WAXS patterns from relatively small, rigid proteins, such as cytochrome *c*. In Fig. 1, the WAXS pattern calculated from CRYSOL is compared to that calculated from atomic coordinates only (ignoring the hydration layer) and to experimental data. The curve generated from CRYSOL has, in this example, been fit to the experimental data through adjustment of the parameters defining the hydration layer. In the calculations presented in this article, we use default parameters from CRYSOL, including those for hydration layer and contrast. In our analysis of ligand-induced structural changes in proteins,[2] we found that differences among patterns generated with CRYSOL accurately reflected the observed differences even though small systematic discrepancies between calculated and observed patterns were observed. In other words, any systematic discrepancies between calculated and observed patterns were uniform for all cases studied. These observations indicate that the use of CRYSOL is entirely appropriate for the systematic study of differences among WAXS patterns as presented here.

Results and Discussion

Number of independent parameters contained in a WAXS pattern

The interatomic distance information contained within a WAXS pattern is defined in the Debye formula that expresses the intensity of solution scattering, $I(s)$, as a function of spacing, s (where $s = 2\sin(\theta)/\lambda$; 2θ is the scattering angle and λ is the wavelength of incident X-rays). In terms of the lengths of the interatomic vectors, r_{ij}, the intensity can be written as,[14]

$$I(s) = \sum I_i(s) + 2\sum\nolimits_{ij} F_i(s)F_j(s)\left(\sin\left(2\pi r_{ij}s\right)/\left(2\pi r_{ij}s\right)\right), \tag{1}$$

where $I_i(s)$ is the scattered intensity due to atom i, $F_i(s)$ is the atomic scattering factor of atom i, and the sum is over all atoms. At small angles of scattering, long interatomic vectors dominate, allowing the use of SAXS for the determination of protein size and shape. At wider angles (i.e., higher s), shorter intermolecular vectors contribute proportionately more to the intensity, incorporating information about protein secondary and tertiary structure into the scattering data.

The pair correlation function, $p(r)$, of a protein can be calculated from the spherical Fourier transform of the measured intensity[15] as

$$p(r) = \int_0^\infty I(s)[\sin(2\pi rs)/2\pi rs]4\pi s^2\mathrm{d}s. \tag{2}$$

$p(r)$ is literally a histogram of all the interatomic vector lengths in the protein. Unnecessary for the calculations carried out in this article, this relationship nonetheless demonstrates the intimate relationship between scattered intensity and the distribution of interatomic distances in the protein.

Because the molecules are tumbling in solution, the directional information associated with each interatomic vector is lost. Limited experimental investigations to explore the amount of structural information in a WAXS pattern have been carried out,[1–4] but no quantitative evaluation of this information has yet been performed. As a first step, an information-theoretic analysis of solution scattering can be made[15,16] analogous to that for one-dimensional[17] or fiber diffraction data.[18,19] This provides a measure of the number of independent parameters that can be determined from the data. The Shannon sampling theorem indicates that a band-limited function can be perfectly reconstructed from samples taken at intervals of one over the band width,[16] and the number of those samples provides a measure of the information contained in the entire, continuous function. For WAXS data, the 'bandwidth' corresponds to twice the maximum spatial extent of the protein or, equivalently, the maximum extent of the pair correlation function. Scattering from a protein with maximum linear dimension of 35 Å can be reconstructed from samples taken at intervals of 1/70 Å^{-1} or less. A WAXS pattern from this 35-Å-diameter protein extending out to a spacing of $s = 0.2$ Å^{-1} contains approximately 70/5 or 14 independent measurements or parameters. To a spacing of 0.5 Å^{-1}, it contains ~35 independent parameters.

Although knowledge of the number of independent parameters defined by a WAXS pattern quantifies the information content, it provides little insight into how useful that information is in distinguishing between protein folds. The amount of information in a WAXS pattern that is relevant to distinguishing among protein structures can be estimated through the analysis of multiple WAXS patterns computed from crystallographic coordinates. To take this approach, we represent a WAXS pattern as a multidimensional vector with components that correspond to the intensities in the pattern. For instance, a pattern extending to $s = 0.2$ Å^{-1} sampled at intervals of 0.0025 Å^{-1} is equivalent to a vector in 80 dimensions. For virtually all proteins, this would constitute oversampling according to Shannon,[16] as adjacent intensities sampled on this grid will not be independent. Nonetheless, this does not alter the results of the analysis, since the number of significant dimensions can be determined by a principal components analysis (PCA) that automatically discards redundant information.[20]

As a surrogate for experimental patterns, a data set suitable for PCA was obtained by using 498 WAXS patterns calculated from atomic coordinates of protein domains selected to represent the broadest possible range of known protein folds.[8] These small domains exhibit typical characteristic dimensions of about ~35 Å, which makes it appropriate to compare the properties of this distribution with the results of the naïve sampling theorem calculation outlined above. A PCA was carried out on the resulting set of 498 vectors representing these WAXS patterns and the corresponding eigenvectors and eigenvalues were obtained. The eigenvector corresponding to the largest eigenvalue represents the direction in WAXS space that most completely distinguishes among the members of the set (in this case, the set of 498 WAXS patterns). Eigenvectors that correspond to very small eigenvalues represent directions in this space that do not distinguish between patterns. Consequently, the distribution of eigenvalue magnitudes provides a measure of the number of independent parameters that distinguish the WAXS patterns generated by these 498 proteins. Figure 2 presents the magnitudes of the largest eigenvalues computed from the 498 WAXS patterns by using intensities that extend out to two different resolution limits: $s=0.2$ and 0.5 Å^{-1}. Eigenvectors corresponding to eigenvalues with magnitude less than one generally provide no information about the distribution of samples in a space constructed in this manner. For the low-resolution case ($s<0.2$ Å^{-1}), the number of significant eigenvectors is approximately 13. This estimate corresponds closely to that made on the basis of the Shannon sampling theorem as indicated previously. Use of data extending out to $s=0.5$ Å^{-1} increases this number to approximately 35, again roughly consistent with the Shannon sampling theorem. This correspondence substantiates the analysis and suggests that all portions of a scattering pattern contribute to distinguishing among WAXS patterns from different proteins. If a region of the WAXS patterns were virtually identical among all proteins (as suggested by Zagrovic and Pande[6]) then we would expect the amount of information as estimated from the PCA to be significantly less than that calculated from a naïve sampling theorem approach.

The nth WAXS pattern, $I_n(s)$, can be expressed as a linear sum of the eigenvectors, $E_i(s)$,

$$I_n(s) = \sum_i e_{ni} E_i(s), \tag{3}$$

with coefficients e_{ni}. The set of e_{ni} constitutes an alternate representation of the information contained within the scattering pattern. The intensity, $I_n(s)$, can be represented as either a continuous function of s or as a multidimensional vector with components e_{ni}. The lowest-order eigenvectors represent the features of the scattering patterns that most distinguish between the 498 scattering patterns used to construct them. Figure 3a contains plots of the five lowest-order eigenvectors. They are dominated by small-angle features, and those features move to progressively wider angles as order increases. Figure 3b contains eigenvectors 5, 10, 15 and 20. As order increases, dominant features shift to wider angles. Nevertheless, even the lowest-order eigenvectors maintain a significant amplitude at wider angles as would be required to explain the effect of resolution on the distribution of proteins in the space defined by these eigenvectors.

Secondary-structure information in a WAXS pattern

Secondary structures such as α-helices and β-sheets are identified by a structural repeat and contain characteristic patterns of interatomic vector lengths. For instance, α-helices have characteristic periodicities of 5.4 Å (pitch) and 10 Å (center-to-center packing distance). β-Sheets are characterized by a 4.7 Å (strand-to-strand) distance. Because WAXS data reflect the interatomic distances in a protein, we determined the degree to which it measures the relative abundances of α-helices and β-sheets in proteins as a first step in quantification of fold information in these data.

A training set of 100 proteins (selected from the 498) representing all known fold classes (e.g., α, β, α/β, and α+β) was chosen, their corresponding WAXS patterns were calculated using CRYSOL,[11] and the position of each pattern in WAXS space was determined. As shown in Eq. (3), the vector representing the nth WAXS pattern, I_n, can be expressed as a linear sum of the eigenvectors, E_i, with coefficients e_{ni}. We assumed that information pertain-

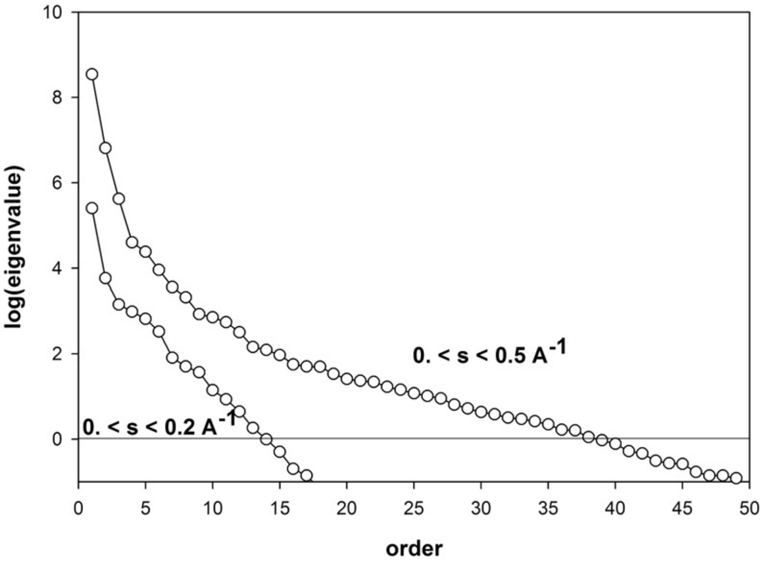

Fig. 2. Plot of the distribution of the magnitudes of eigenvalues computed for the lowest-order eigenvectors in WAXS space *versus* order. When data to s ~0.2 Å^{-1} were used, 13 eigenvalues have a value above 1.0. When data to s ~0.5 Å^{-1} were used, over 35 eigenvalues have a magnitude greater than 1.0.

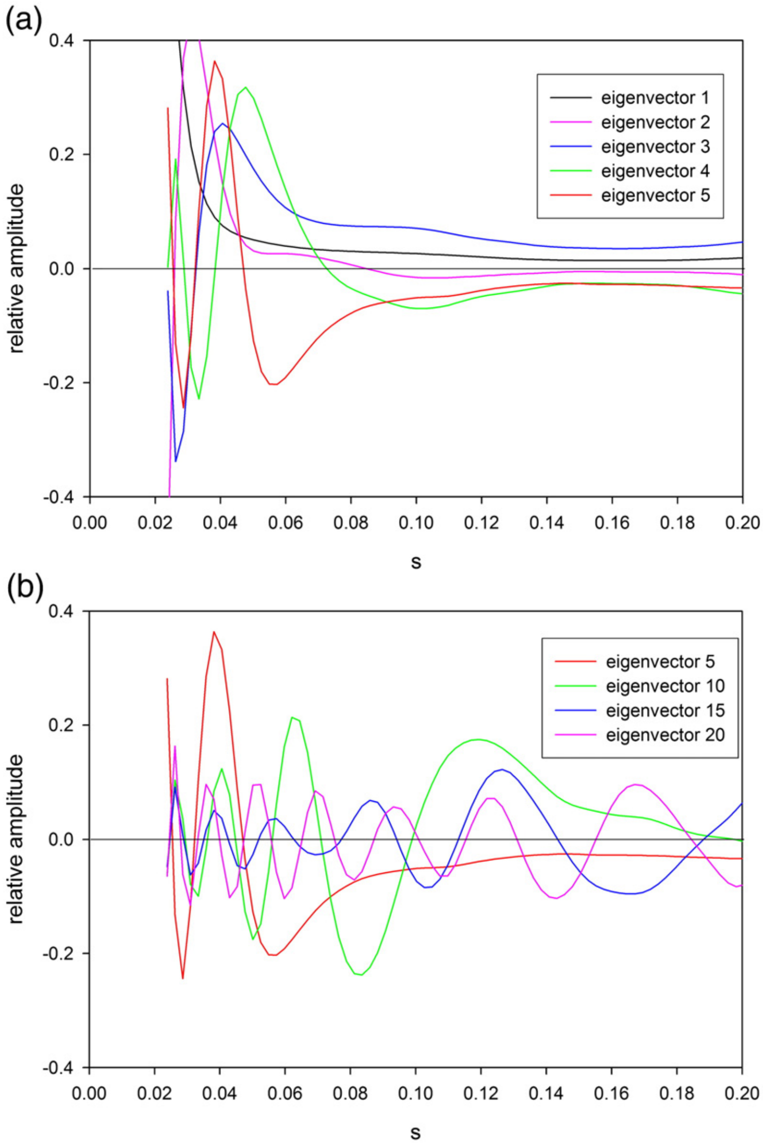

Fig. 3. Relative amplitudes of the WAXS space eigenvectors as a function of s ($\sim 1/d$). (a) Eigenvectors 1–5. (b) Eigenvectors 5, 10, 15, and 20. The lowest-order eigenvectors have dominant features at low spacings. Higher-order eigenvectors have features at progressively larger spacings.

ing to the α-helical content, α_n, of the nth protein is embedded within these coefficients (i.e., within the scattering patterns) and that it can be represented as a linear combination of these coefficients:

$$\alpha_n = \sum\nolimits_i a_i e_{ni}. \tag{4}$$

The parameters, a_i, required to calculate the α-helical content from the WAXS patterns, are assumed to be the same for every protein. The coefficients, e_{ni}, can be calculated from the WAXS pattern. Initial calculations were performed using intensities extending over the interval $0.05<s<0.45$ Å^{-1} and involved use of 30 coefficients, e_{ni}, which, from Shannon sampling considerations correspond to the number of independent parameters expected within this interval. The corresponding 30 unknown a_i values were estimated on the basis of the training set of 100 WAXS patterns calculated from crystallographic coordinates. Since the α-helix and β-sheet contents of these proteins are known from their crystal structures, Eq. (4) corresponds, in this case, to a set of 100 linear equations in 30 unknowns and the unknown a_i can be estimated by the solution of this equation set. Using the a_i calculated from this equation set, we estimated the α-helix and β-sheet contents of the 398 proteins remaining in the test set after removal of the 100-protein training set by using Eq. (4). These estimates were compared with the values determined from their crystallographic coordinates and the results are shown in Fig. 4. The average error in the estimate of α-helical content was 11% and the error in estimation of β-sheet content was 9%. Given the crudeness of the assumptions used to make these calculations, the correlation is quite good. It is possible that more accurate estimates can be made by adding nonlinear terms to Eq. (4).

Our estimation of secondary-structure content from WAXS demonstrates that the patterns of interatomic distances intrinsic to different secondary structures are, indeed, embedded within the WAXS data. We do not, however, believe that WAXS will become a standard method for estimation of secondary structures. CD is widely used for this purpose and has a long history as

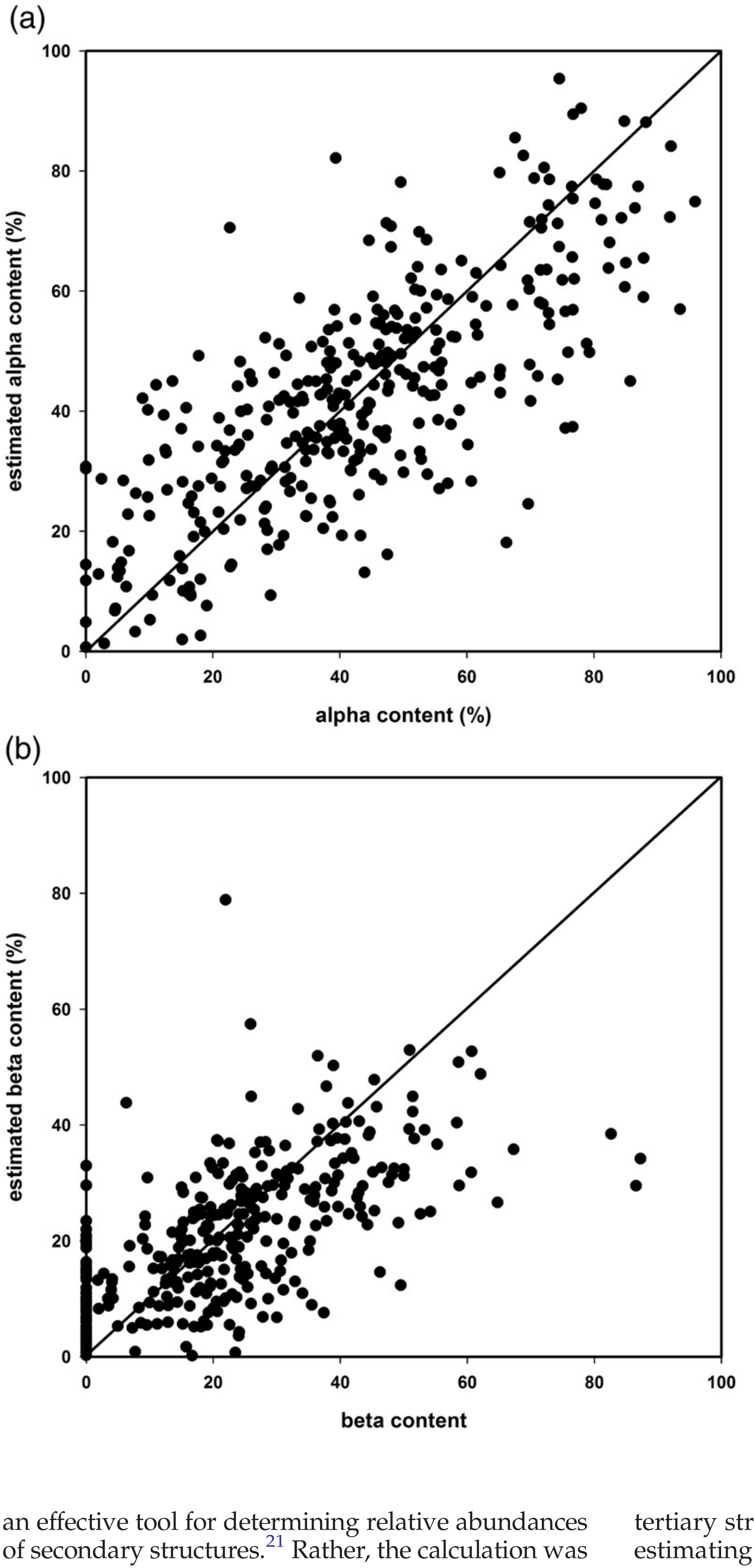

Fig. 4. α-Helix and β-sheet content for 400 proteins of known structure as estimated from WAXS data. (a) Comparison of the estimated and calculated α helix content. (b) Comparison of the estimated and calculated β-sheet content.

an effective tool for determining relative abundances of secondary structures.[21] Rather, the calculation was carried out in order to demonstrate that such information is contained within WAXS data.

Tertiary-structure information in a WAXS pattern (protein fold classification)

Quantitation of the amount of information in a WAXS pattern that is relevant to determining the tertiary structure of a protein is more complex than estimating the abundances of secondary structures. To approach this problem, we carried out three sets of calculations: (a) we examined the distribution of proteins belonging to the super families α, β, α/β and α+β across WAXS space as a function of the resolution of WAXS data used; (b) we compared distances in WAXS space to the *Z*-scores used as a measure of fold similarity in DALI; and (c) we examined the volume of WAXS space that a single

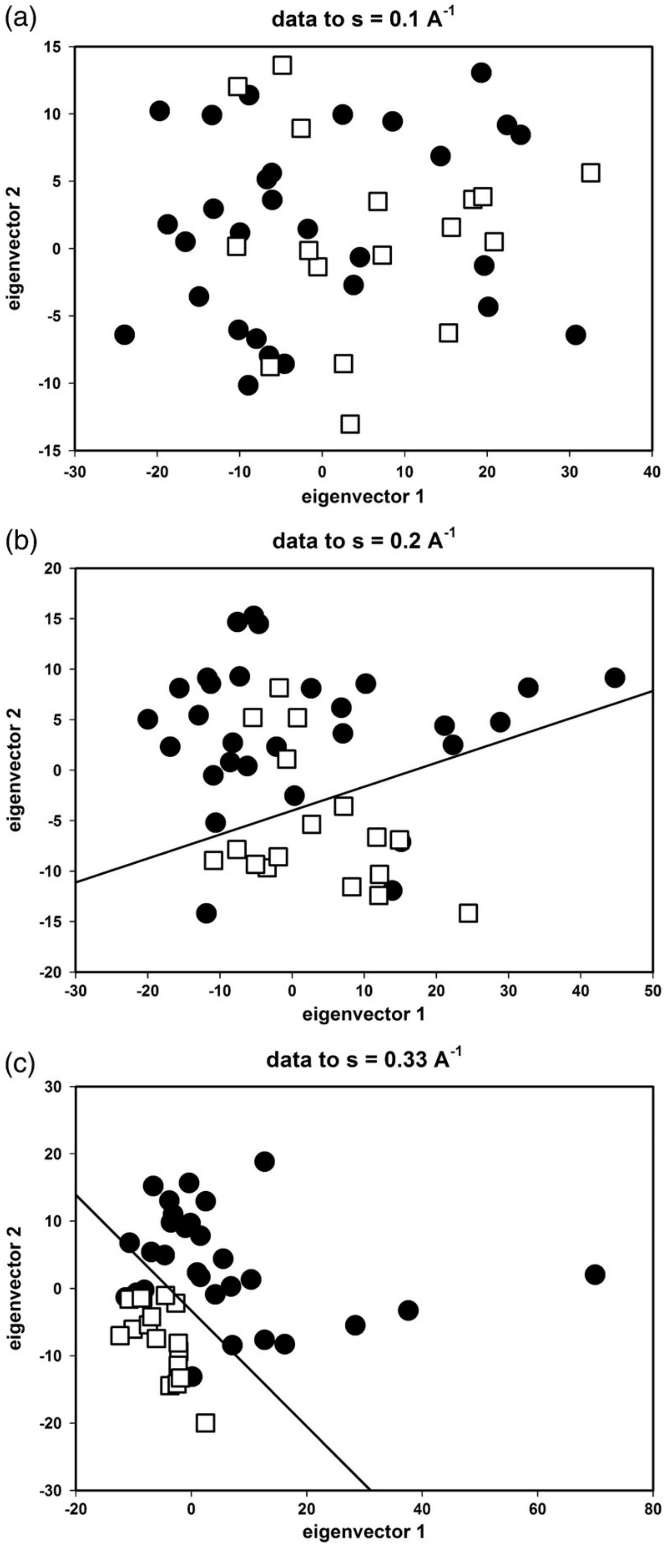

Fig. 5. Distribution of α-helical proteins (circles) and β-sheet proteins (squares) in WAXS space as calculated using data of different resolutions when calculated using data to (top) 10 Å spacing; (middle) 5 Å spacing; and (bottom) 3.0 Å spacing.

fold family occupies and compared it to the distances between members of distinct fold families in that space. The *Z*-score is a measure of the weighted sum of similarities of corresponding intramolecular distances in two proteins as defined by Holm and Sander.[7] We conclude that adjacent fold families overlap in WAXS space, but that the overlaps are confined to families that are both qualitatively and quantitatively similar in structure. This result suggests that on the basis of WAXS data alone, the fold of a protein of unknown structure can be limited to a relatively short list of possibilities by the process of comparing its WAXS pattern with a sufficiently large set of WAXS patterns from proteins of known structure.

We first addressed the distribution in WAXS space of proteins belonging to the four superfamilies, α, β, α/β and α+β, as a function of the resolution of the WAXS data used. Positions in WAXS space were calculated for a training set of 100 proteins (out of the 498) categorized as α, β, α/β or α+β according to SCOP.[8,10] Their distribution in this multidimensional space was visualized using two-dimensional projections onto the directions of the two most significant eigenvectors. In Fig. 5, squares correspond to all β-sheet proteins and circles to all α-helix proteins. When this calculation is carried out using WAXS data to 10 Å spacing ($s \leq 0.1$ Å^{-1}), the distributions of α and β proteins are largely overlapping in this two-dimensional projection. When data to 5 Å spacing are used ($s \leq 0.2$ Å^{-1}), the two distributions are separated with relatively little overlap. When data to 3.0 Å spacing ($s \leq 0.33$ Å^{-1}—corresponding to ~20 independent parameters) are used, the two distributions are almost entirely separate. These results demonstrate that in WAXS patterns at spacings below 0.1 Å^{-1}, little information about secondary or tertiary structure is present, but at higher scattering angles, the presence of specific secondary or tertiary structures appears to be reflected in the data.

A more informative representation of the degree of separation of different fold classes in WAXS space can be obtained with a three-dimensional representation using axes corresponding to three eigenvectors. Many three-dimensional representations are possible, and alternative representations using different combinations of eigenvectors provide distinct views with some choices better visualizing separation of fold families than others. Figure 6 contains a stereo pair showing the distribution of α, α+β, α/β and β-proteins using eigenvectors 2, 3 and 4 (generated using data to 3 Å spacing; $s = 0.33$ Å^{-1}). Figure 7 contains two representative stereo pairs that demonstrate the separation of α-helix proteins from α+β proteins and the separation of β proteins from α/β proteins, respectively. These figures represent separations with only three of the ~20 parameters that can be extracted from WAXS patterns to this resolution. Furthermore, although data to 0.33 Å^{-1} spacing were incorporated, the eigenvectors used here are dominated by relatively low resolution features (see Fig. 3). Complete separation on the basis of these three parameters would not be anticipated. Separation of fold classes appears superior in the three-dimensional figures than in the two-dimensional distributions calculated at moderate resolution (Fig. 5) because the boundaries between fold classes are not parallel to a principal axis and the three-dimensional representation allows some flexibility in choosing the angle of view.

Hou *et al.*[8] noted overlap of protein classes in their analysis of fold-space on the basis of a measure of structural similarity calculated by DALI.[7] Although they suggested that part of the overlap they observed was due to ambiguities in fold designations in SCOP, another possibility is the limited dimensionality of their presentations. Our results indicate that proteins in different fold families are better segregated in three-dimensional representation of WAXS space than in a two-dimensional represent-

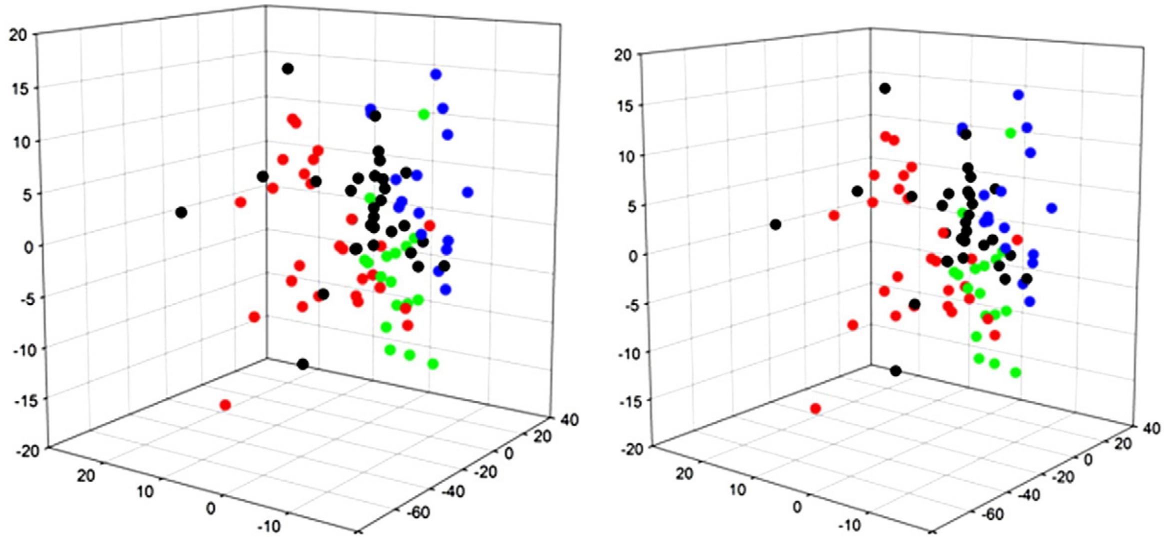

Fig. 6. Stereo pair showing the distribution of α (red); β (blue); α+β (black) and α/β (green) proteins in WAXS space as judged from coordinates 2; 3 and 4 in WAXS space.

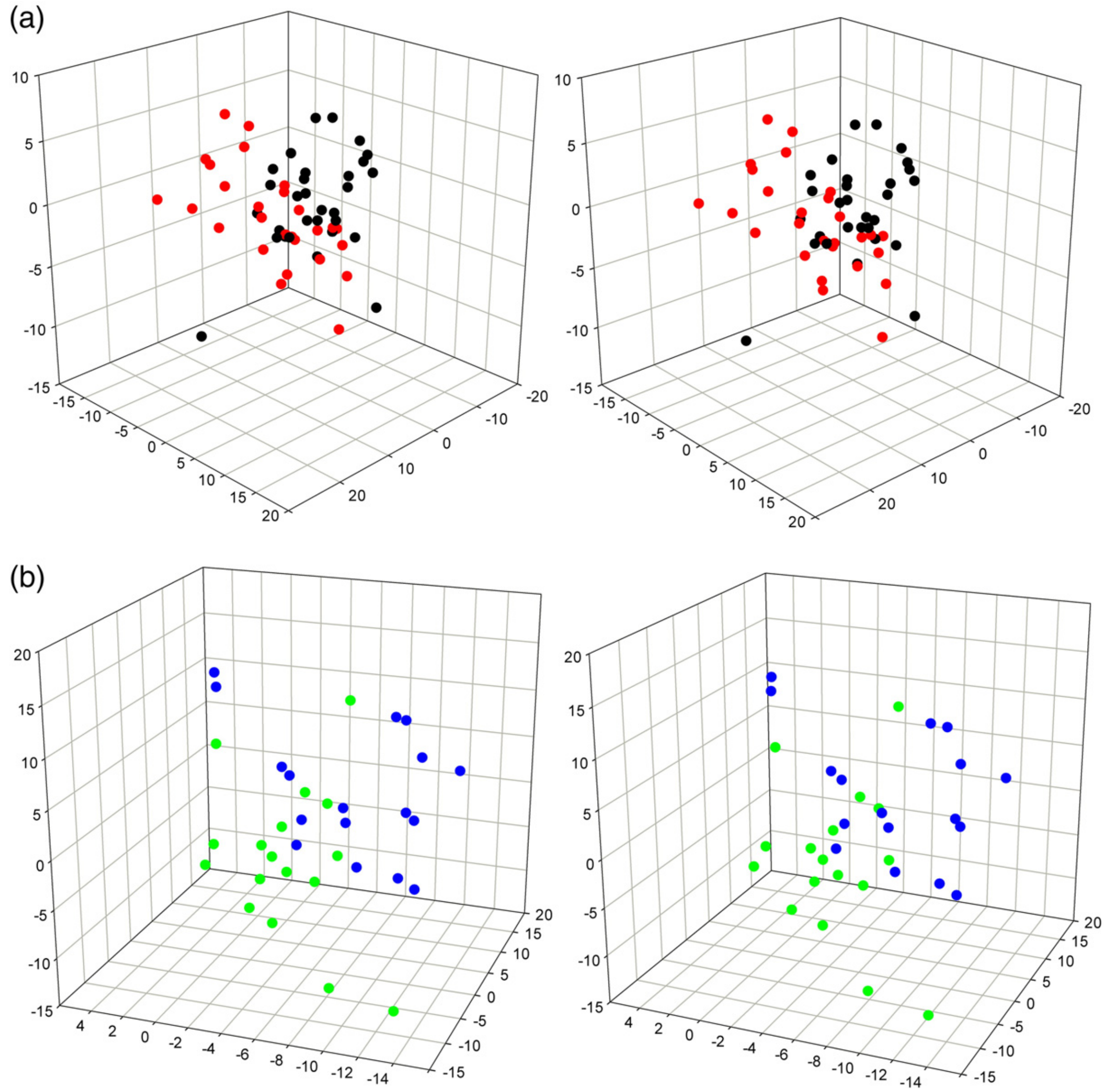

Fig. 7. Stereo pairs showing the distribution of (a) α (red); and α+β (black) and (b) β (blue); and α/β (green) proteins in WAXS space as judged from coordinates 2, 3 and 4 in WAXS space.

ation. Presentation of the distributions computed by Hou *et al.*[8] in higher dimensionality might provide evidence for more complete separation of the superfamilies.

Relative distances in WAXS space *versus* fold space

It may be possible to assign the fold of a protein of unknown function via comparison of its WAXS pattern with those of proteins of known structure if two conditions are met: (a) proteins that give rise to very different WAXS patterns have very different structures and (b) proteins giving rise to similar WAXS patterns have similar structures. These two conditions were tested with a data set of proteins chosen to include both close structural homologs and structurally unrelated pairs. Both the *R*-factor between WAXS patterns and the *Z*-score (computed using DALI[7]) were calculated for all pairs of proteins in this set as metrics for relatedness in WAXS pattern and structure, respectively. The *R*-factor used for this comparison was $\int|I_1(s)-I_2(s)|\mathrm{d}s/\int I_n(s)\mathrm{d}s$ where all scattering patterns, n, were normalized by setting $\int I_n(s)\mathrm{d}s$ equal for all n. Note that distance in WAXS space corresponds to differences in the linear coefficients of eigenvectors required to reconstruct a WAXS pattern, which in turn correspond to differences in relative intensities. Since we have normalized the WAXS patterns, the unit of distance in WAXS space is dimensionless.

Figure 8 shows that (i) pairs of proteins recognized as close homologues by a high *Z*-score in DALI had low *R*-factors between their WAXS patterns and (ii) pairs of proteins with high *R*-factors (i.e., very different WAXS patterns) had low *Z*-scores (i.e., very different folds). These results indicate that distance in WAXS space is roughly correlated with distance

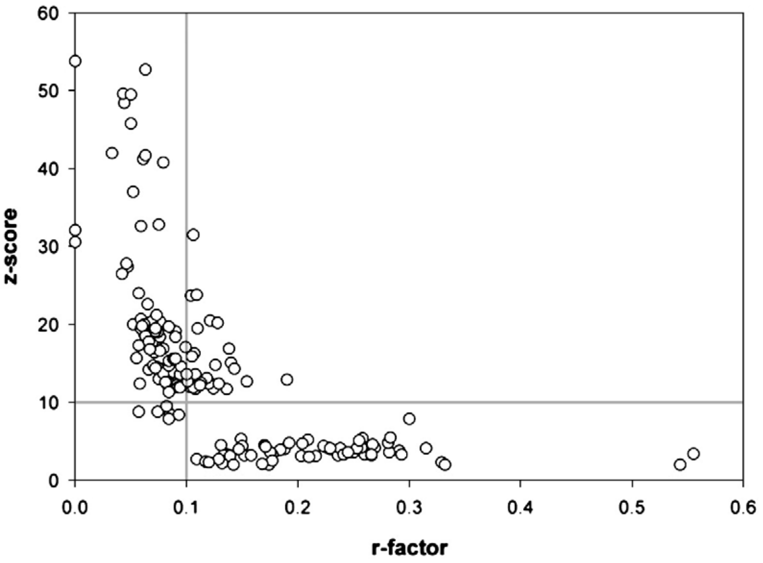

Fig. 8. The correspondence of Z-score as calculated by DALI and R-factor as calculated from WAXS patterns for all possible pairings of 14 proteins chosen to represent close structural homologs (high Z-score; low R-factor) and unrelated structures (low Z-score; high R-factor).

in fold space with DALI Z-scores as a parallel but distinct metric. Over short distances, this correlation is very good—high Z-scores correspond to low R-factors. Over longer distances, the correlation is weaker due to the different metrics defining structural differences among proteins by using WAXS *versus* DALI. For the successful use of WAXS in assigning protein fold, only the correlation over short distances is required. The fact that DALI and WAXS provide different measures of the distances among very different structures is unimportant as long as they both recognize that relationship as 'distant'.

Distribution of fold families in WAXS space

Extrapolating from this test set, the distribution of representative folds in WAXS space appears to place those folds that are most similar (as judged by DALI) close to one another, and those folds that are very different much further apart. Although this is consistent with the hypothesis that WAXS may be valuable for assignment of protein folds based on comparisons with WAXS patterns from proteins of known fold, we need to determine what the error bars are on a fold assignment made on the basis of WAXS data. How much volume does a single fold family occupy in WAXS space? To what extent do similar fold families overlap? To answer these questions, we looked at the distribution of members of specific fold families relative to one another and to proteins from other fold families. When data to 0.33 Å^{-1} are used (~20 significant dimensions), the average distance in WAXS space between proteins with folds distinct from one another is 23.0 and the average distance to the nearest protein that exhibits a distinctly different fold is 9.5. Again, these numbers are dimensionless and take on meaning only when their relative magnitudes are compared to one another. The average distance between a protein that belongs to the myoglobin fold family and other members of the myoglobin family ranges from 2.9 to 4.3 depending on whether the protein is central or peripheral to the family. Proteins representing only three of the 498 distinct folds analyzed have an average distance of less than 5.0 to members of the myoglobin fold family. These are 1hw1, 1al0 and 1elk. The regions occupied by the fold families represented by these three structures overlap that of the myoglobin fold family in WAXS space. As shown in Fig. 9, each of these structures exhibits a fold that includes layers of alpha helical structure arranged in a manner similar to that of myoglobin. On the basis of WAXS data alone, a protein of unknown structure belonging to the myoglobin fold family might be identified as belonging to the myoglobin family or to one of these three distinct but qualitatively similar fold families. This relatively low level of ambiguity is representative of the uncertainty that would be intrinsic to assignment of folds on the basis of WAXS data alone.

A larger, more diverse fold family such as the triose phosphate isomerase (TIM) barrel occupies considerably more volume in WAXS space. Based on 50 members exhibiting less than 90% sequence homology to one another, the average distance between TIM family members is about 15—roughly 50% greater than the average nearest-neighbor distance between distinct folds in WAXS space. Consequently, 5 to 10 protein folds may exhibit WAXS patterns sufficiently similar to that of a TIM barrel protein that they could not be distinguished from it on the basis of WAXS data alone.

Examination of eight fold families in WAXS space suggests that the results for the myoglobin and TIM barrel families are typical. Although fold families that occupy adjacent regions of WAXS space appear to overlap, the degree of overlap is limited to adjacent families with similar folds. The fold families examined were those representing close structural homologues (as judged by DALI) to 2tbd (SV40 T-antigen), 1ayl (phospho*enol*pyruvate carboxykinase), 1cao (carbonic anhydrase), 1c9ja (subtilisin), 1a2y (lysozyme), 1jw8a (myoglobin), 1g7u

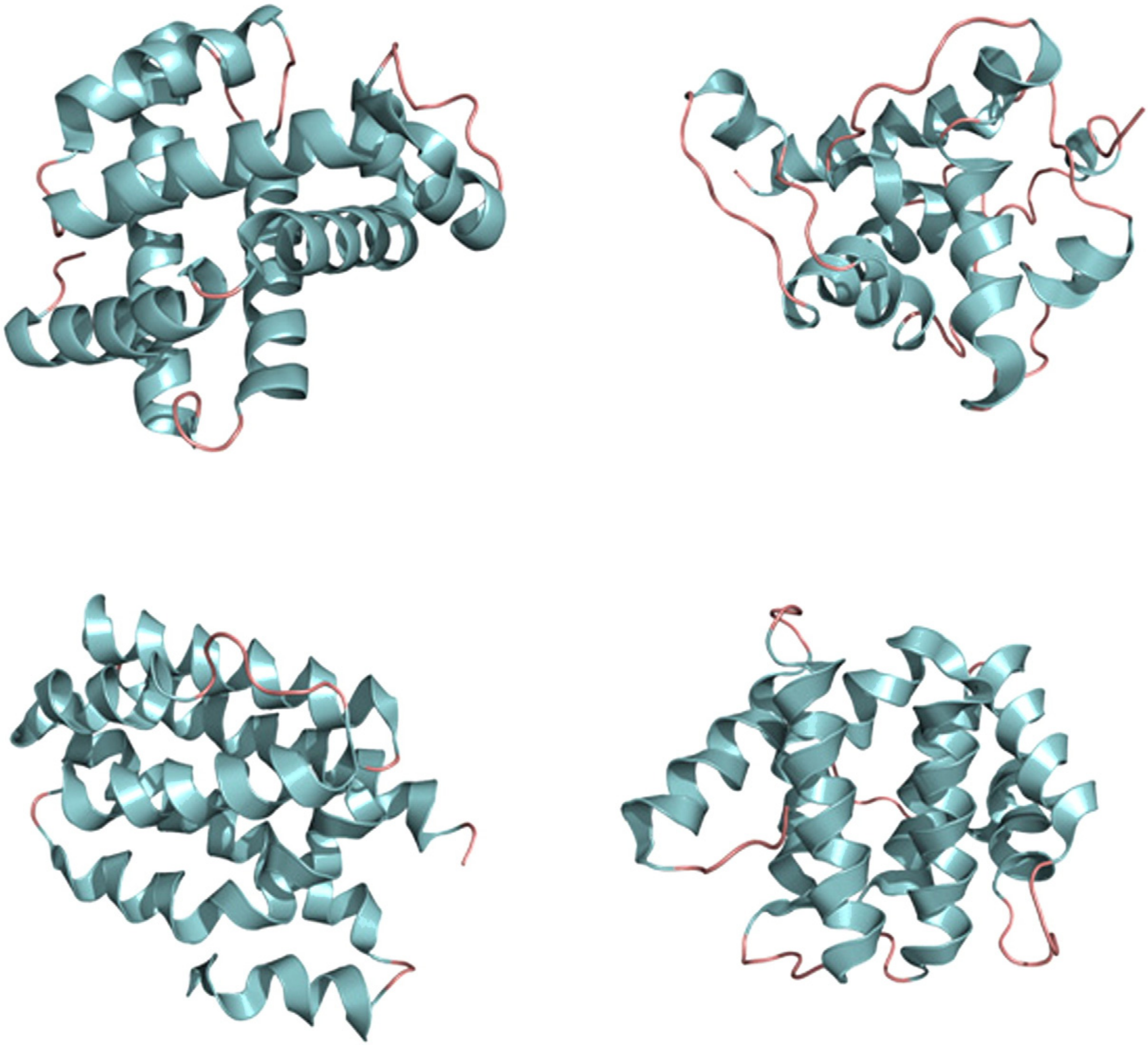

Fig. 9. The structures of myoglobin (top left) and three proteins (1hw1, 1al0, and 1elka representing fold families that occupy regions of WAXS space that overlap the region occupied by the myoglobin fold family.

(2-dehydro-3-deoxyphosphooctonate aldolase) and 1qtta (product of mtcp1 oncogene). These analyses indicate that with a sufficiently large reference set of WAXS patterns, a WAXS pattern from a protein of unknown structure should be sufficient to construct a short list of 2 to 10 fold families of which the protein could be a member. With this short list as a starting point, additional informatics analysis such as sequence comparison with proteins from the candidate fold families could well result in a unique designation.

Multidomain proteins

The above analyses have been carried out with single domains of proteins. What is the impact of multiple domains on the use of WAXS for structure assignment? The space of all possible protein structures is considerably more complex than the space of all possible domain structures.[9] Here, we consider a single example of hemoglobin and myoglobin as representative of the kind of behavior that multidomain proteins may exhibit. Hemoglobin is made up of four subunits, each homologous in structure to myoglobin. Although the distribution of interatomic distances in hemoglobin is likely to have considerable similarity to that of myoglobin, hemoglobin will have many long interatomic vectors that do not occur in myoglobin. How will that impact its placement in WAXS space? Since low-order eigenvectors include information about the size and shape of a molecule, the coefficients e_i for hemoglobin will presumably be different from those for myoglobin for low-order eigenvectors, but much more similar for higher-order eigenvectors that contain more information about shorter interatomic vectors. As a demonstration, consider the comparison between hemoglobin and the small, similar proteins, myoglobin and sperm lysin. Their structures are depicted in Fig. 10a, with the differences between their coefficients, e_i, plotted in Fig. 10b. For low-order eigenfunctions, hemoglobin coefficients are actually closer than myoglobin to sperm lysin. This is due to the larger spatial extent of lysin

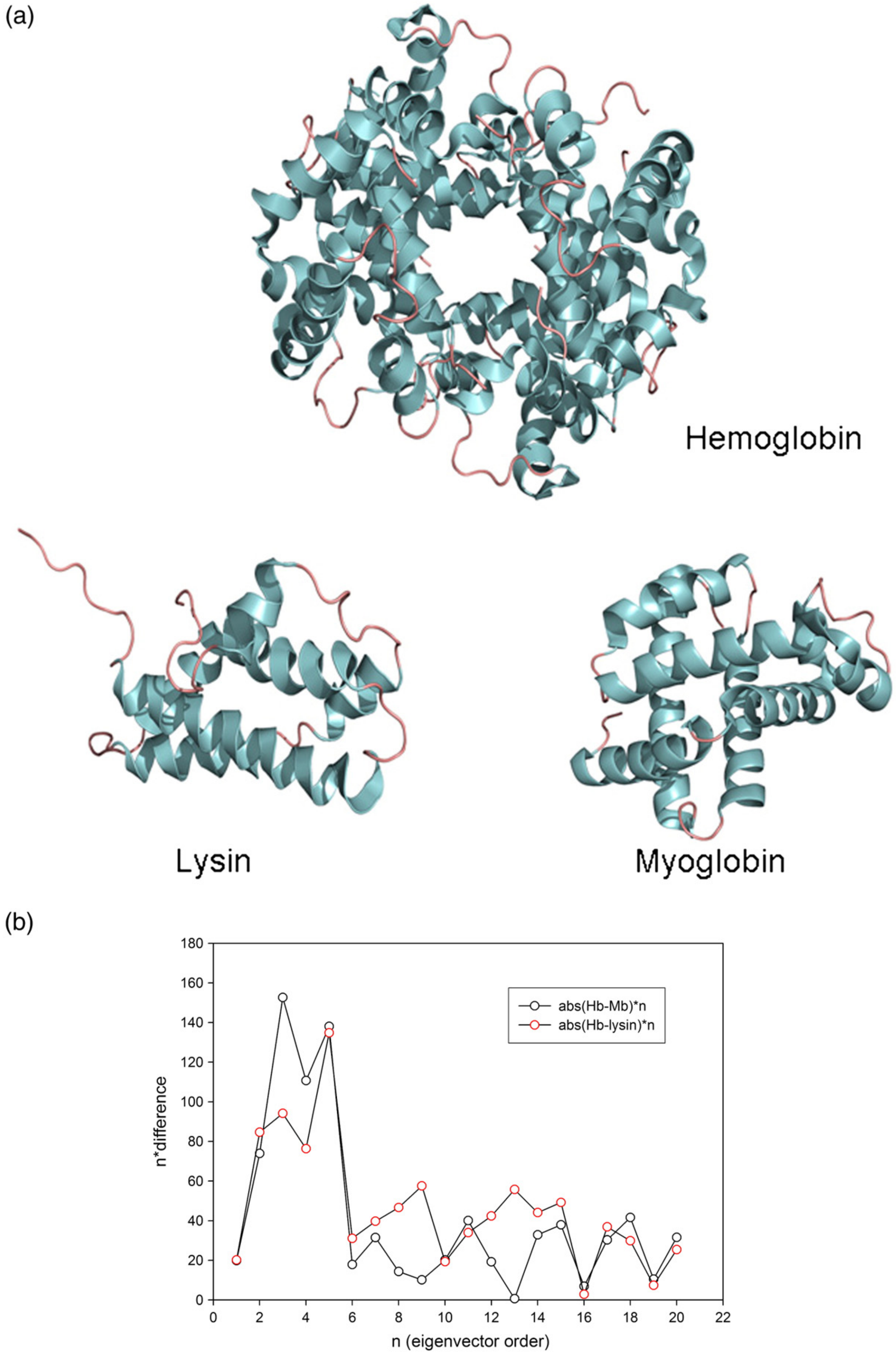

Fig. 10 (*legend on next page*)

relative to myoglobin. For higher-order eigenfunctions, hemoglobin is closer to myoglobin, as one might expect given the similar distributions of short interatomic vectors in the two proteins. Although the uses of WAXS for characterization of multidomain proteins will be more complex than for single-domain proteins, these results suggest that there is considerable relevant information in WAXS patterns.

Prediction of WAXS patterns from atomic coordinates

The utility of WAXS is due, in part, to the fact that WAXS patterns can be calculated from atomic coordinates. The WAXS patterns used in this study were computationally generated using CRYSOL, the most widely used program for calculation of solution scattering patterns from atomic coordinates.[11] This program, originally designed for SAXS data, has some limitations when extended to the calculation of solution scattering to wider angles of scatter. For small proteins—such as the individual domains considered here—these limitations are likely to be relatively modest and should not impact the conclusions of this study. See Fig. 1 for an example. For calculation of wider angle scattering from larger proteins, the approximations implicit in the mathematical representation of electron density in CRYSOL and its treatment of excluded volume and solvation layer may lead to systematic errors. Given this limitation, the power of WAXS as a method for structural characterization would be greatly enhanced by a database of experimental WAXS patterns from proteins of known structure. This type of database has been previously used with great success in analysis of CD spectra.[13] As we have shown here, a database containing WAXS patterns from representative members of all known fold families would make possible construction of a short list of possible folds for a protein of unknown structure. In principle, it should be possible to construct that database computationally from atomic coordinate sets in the Protein Data Bank (PDB). In practice, there remain a number of barriers to doing so including limitations of available software, the effect of intramolecular motions on observed scattering,[22] and possible systematic errors in the data collected. Progress in all of these areas is expected over the next few years.

The calculations outlined here indicate that WAXS has the potential for assigning a short list of possible folds to a protein of unknown structure on the basis of a simple experiment to collect a WAXS pattern and a comparison of its WAXS pattern with those of proteins of known structure. As such, it could provide an important adjunct to other structural and biophysical methods for characterization of proteins of unknown structure and function.

Experimental data

Construction of a large database of experimental WAXS patterns to act as a basis set for comparisons could provide a foundation for the structural classification of proteins of unknown structure. An attractive alternative—the use of a computationally generated basis set for classification of experimental data—could not, at this time, result in accurate predictions. This, in spite of the relatively close correspondence between observed WAXS patterns and those calculated by CRYSOL (see, for instance, Fig. 1). Experimental WAXS data from proteins can be measured to at least 2 Å spacing. The high stability and brilliance of third-generation synchrotron sources[1,2,22] make possible the required subtraction of background scattering from specimen chamber (capillary) and buffer even though this subtraction requires knowledge of the volume of buffer excluded by the presence of protein.[22] However, experimental patterns are influenced by temperature, protein concentration and structural fluctuations[22] and we do not yet know the degree to which details of the hydration layer may vary from protein to protein. CRYSOL, originally designed for calculation of SAXS patterns, is not designed to take all of these factors into account. The results outlined here indicate that a program capable of accurate calculation of WAXS patterns could have a very substantial impact.

Methods

PDB files

The 498 PDB files representative of all classes of known folds were downloaded from the Research Collaboratory for Structural Bioinformatics (taken from SCOP release 1.55) using the list provided by Dr. Sung Ho Kim and were identical to those used by his group to define and characterize fold space.[8] Parent files were modified to delete portions of the protein that were not included within that SCOP fold class (i.e., multidomain proteins). A list of the 498 domains used is provided as supplemental material.

Calculation of WAXS patterns from atomic coordinates

WAXS patterns were calculated from the PDB files with the program CRYSOL[11] modified to accommodate up to 50 spherical harmonics, thereby improving the accuracy of the computation at wide angles. Default values were used

Fig. 10. (a) Structure of hemoglobin; myoglobin and sperm lysin. (b) Differences in the coefficients of the vectors representing WAXS patterns of hemoglobin, myoglobin and sperm lysin. For the lowest orders, the coefficients of hemoglobin are closer to those of lysin than to myoglobin. For higher orders, hemoglobin is closer to myoglobin than to lysin. The coefficients have been multiplied by the order of the eigenvector to represent more clearly the differences at higher order.

for electron density of the solvent, contrast of the solvation shell, average atomic radius and excluded volume.

Construction of WAXS space

PCA, including calculation of eigenvectors and eigenvalues, used standard algorithms and codes from Numerical Recipes in FORTRAN.[23] Additional code for calculating positions of individual WAXS patterns in WAXS space and estimation of secondary structural elements was written in FORTRAN and are available from the authors on request.

Acknowledgements

We thank Dr. Sung Ho Kim for the list of PDB files used in this analysis, Dr. Sanghyun Park for carrying out the calculations presented in Fig. 1, and Drs. Park and Jay Bardhan for enlightening discussions. This work was supported by the U.S. Department of Energy under contract DE-AC-02-06CH11357.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.08.038

References

1. Fischetti, R. F., Rodi, D. J., Mirza, A., Kondrashkina, E., Irving, T. & Makowski, L. (2003). Wide angle x-ray scattering of proteins: effect of beam exposure on protein integrity. *J. Synchrotron Res.* **10**, 398–404.
2. Fischetti, R. F., Rodi, D. J., Gore, D. B. & Makowski, L. (2004). Wide angle x-ray solution scattering as a probe of ligand-induced conformational changes in proteins. *Chem. Biol.* **11**, 1431–1443.
3. Hirai, M., Iwase, H., Hayakawa, T., Miura, K. & Inoue, K. (2002). Structural hierarchy of several proteins observed by wide-angle solution scattering. *J. Synchotron Radiat.* **9**, 202–205.
4. Hirai, M., Koizumi, M., Hayakawa, T., Takahashi, H., Abe, S., Hirai, H. *et al.* (2004). Hierarchical map of protein unfolding and refolding at thermal equilibrium revealed by wide-angle x-ray scattering. *Biochemistry*, **43**, 9036–9049.
5. Tiede, D. M., Zhang, R. & Seifert, S. (2002). Protein conformations explored by difference high-angle solution X-ray scattering: oxidation state and temperature dependent changes in cytochrome *C. Biochemistry*, **28**, 6605–6614.
6. Zagrovic, B. & Pande, V. S. (2006). Simulated unfolded-state ensemble and the experimental NMR structures of villin headpiece yield similar wide-angle solution X-ray scattering profiles. *J. Am. Chem. Soc.* **128**, 11742–11743.
7. Holm, L. & Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins: Struct., Funct., Genet.* **33**, 88–96.
8. Hou, J., Sims, G. E., Zhang, C. & Kim, S. H. (2003). A global representation of the protein fold space. *Proc. Natl Acad. Sci. USA*, **100**, 2386–2390.
9. Hou, J., Jun, S. R., Zhang, C. & Kim, S. H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl Acad. Sci. USA*, **102**, 5641–5642.
10. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database fo the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
11. Svergun, D., Barberato, C. & Koch, M. H. J. (1995). CRYSOL—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773.
12. Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *J. Appl. Crystallogr.* **11**, 693–694.
13. Lees, J. G., Miles, A. J., Wien, F. & Wallace, B. A. (2007). A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics*, **22**, 1955–1962.
14. Guinier, A. (1994). X-ray Diffraction in Crystals, Imperfect Crystals and Amorphous Bodies, pp. 49, Dover Publications, Mineola, NY.
15. Luzzati, V. & Tardieu, A. (1980). Recent developments in solution x-ray scattering. *Annu. Rev. Biophys. Bioeng.* **9**, 1–29.
16. Shannon, C. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423; 623–656.
17. Makowski, L. (1981). The use of continuous diffraction data as a phase constraint: I. One-dimensional theory. *J. Appl. Crystallogr.* **14**, 160–168.
18. Makowski, L. (1982). The use of continuous diffraction data as a phase constraint: II. Application to fiber diffraction data. *J. Appl. Crystallogr.* **15**, 546–557.
19. Makowski, L. (1991). An estimate of the total number of independent structural parameters measurable in a fiber diffraction pattern. *Acta Crystallogr. A*, **47**, 562–567.
20. Lebart, L., Morineau, A. & Warwick, K. M. (1984). Multivariate Descriptive Statistical Analysis Wiley, New York.
21. Wallace, B. A. (2000). Synchrotron radiation circular-dichroism spectroscopy as a tool for investigating protein structures. *J. Synchrotron Res.* **7**, 289–295.
22. Makowski, L., Rodi, D. J., Mandava, S., Minh, D., Gore, D. & Fischetti, R. F. (2008). Molecular crowding inhibits intramolecular breathing motions in proteins. *J. Mol. Biol.* **375**, 529–546.
23. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1993). Numerical recipes in FORTRAN: The art of scientific computing Cambridge University Press, New York.